# Business Understanding

## Background and Overview

The project involves the development of a predictive model for customer churn using the available dataset, aiming to identify customers at risk of leaving the service. This project directly concerns stakeholders such as the marketing and sales teams, customer service departments, and upper management, with the potential to significantly impact customer retention and overall profitability. While the primary data source is the provided customer dataset, the project's scope encompasses the development and validation of the predictive model, along with the identification of key features influencing churn. Stakeholders' understanding and alignment regarding the project's objectives, scope, and expected outcomes are crucial, as clear communication is vital to ensure a unified understanding across different parts of the organization.

## Business problem

SyriaTel, a leading telecommunications company, is facing challenges with customer churn, where customers discontinue their services with the company. Customer churn not only results in revenue loss but also impacts the company's reputation and market competitiveness. To mitigate this issue, SyriaTel aims to identify predictive patterns and develop a robust classifier to forecast whether a customer is likely to churn in the near future.

## Objectives

This project aims:

1. To develop a binary classification, model that forecasts if a client will "soon" terminate their relationship with SyriaTel.
2. To determine what factors influence customer churn
3. To determine the best model for predicting customer churn
4. To evaluate how insights from feature importance can help improve customer churn

## Significance

By accurately identifying customers at risk of churn, the company can proactively implement retention strategies to mitigate churn and enhance customer loyalty.

### Research Questions

The project aims at answering the following questions:

- What were the factors influencing customer churn?
- What is the best model for predicting customer churn?

- How can the insights from feature importance help improve customer churn?

## Data Understanding

The data is from Syrian Tel Communication company retrieved from Kaggle (link: https://www.kaggle.com/datasets/becksddf/churn-in-telecoms-dataset/data). The structure of the data included 21 columns and 3333 rows. The columns detail the different kinds of attributes (refer to the summary section below) while the rows represent customers recorded in the dataset. The dataset contains continuous and categorical variables. The target variable used is churn and the rest of the variables served as predictors except for state and phone number.

Here's a brief summary of each column:

| |
|---|
| State: Customer's state. |
| Account Length: Duration of the customer's account. |
| Area Code: Telephone area code. |
| Phone Number: Unique phone number. |
| International Plan: Whether the customer has an international calling plan. |
| Voice Mail Plan: Whether the customer has a voicemail plan. |
| Number Vmail Messages: Number of voicemail messages. |
| Total Day Minutes/Calls/Charge: Total daytime call duration, count, and charges. |
| Total Eve Minutes/Calls/Charge: Total evening |

## Data Preparation

During this stage:

i. The data is observed to have no missing values and duplicates.
ii. Categorical data was transformed to numerical data with label encoder.
iii. Normalizing numeric data using Min-Max Scaler.

## Modelling

Modelling considerations:

1. Task Type: This project involves classification to predict customer churn.

2. Models: We'll experiment with logistic regression, decision trees, random forests, and gradient boosting.
3.Class Imbalance Management: We'll address overfitting using SMOTE and hyper parameter tuning.
4. Regularization: We'll use regularization techniques like L1 or L2 regularization to prevent overfitting.
5. Validation: We'll employ k-fold cross-validation to ensure our models generalize well.
6. Loss Functions: We'll use entropy loss and gini for model training.
7. Performance Threshold: Success will be determined by achieving predefined performance metrics thresholds. These considerations guide our model selection, training, and evaluation process, ensuring effective churn prediction.

### Evaluation

The metrics used to evaluate models:

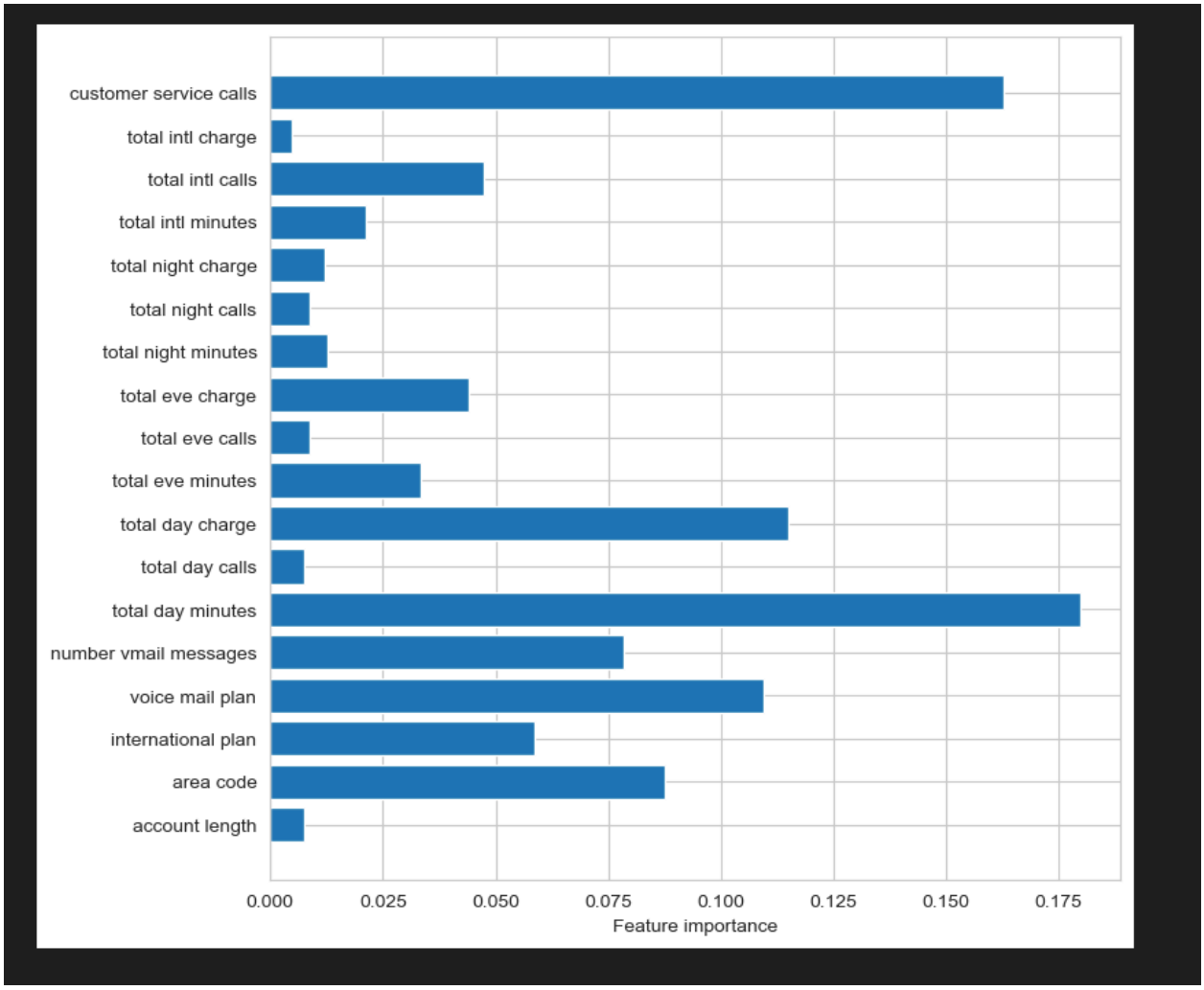
a. Accuracy score
b. Area under curve (auc)

Best model is selected best on high performance .

## Findings of EDA

### Feature Importance

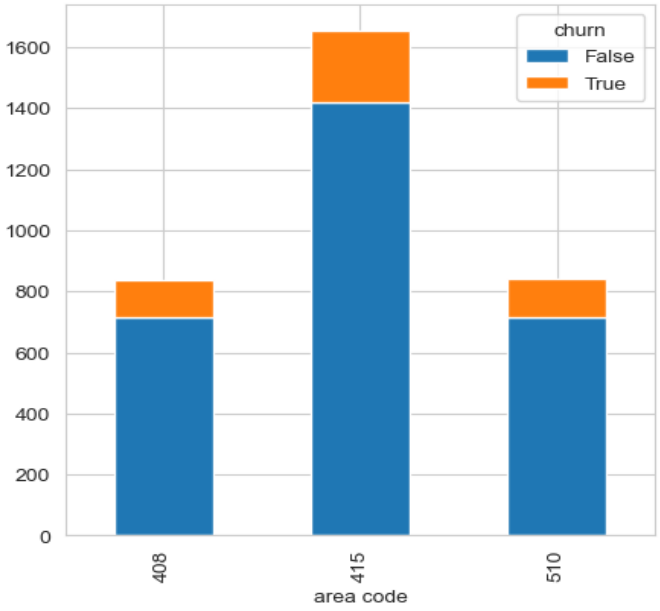
Feature importance analysis helps identify which features have the most influence on predicting churn. By knowing which factors contribute the most to churn, the company can prioritize them in its retention strategies and focus resources on addressing those factors effectively.



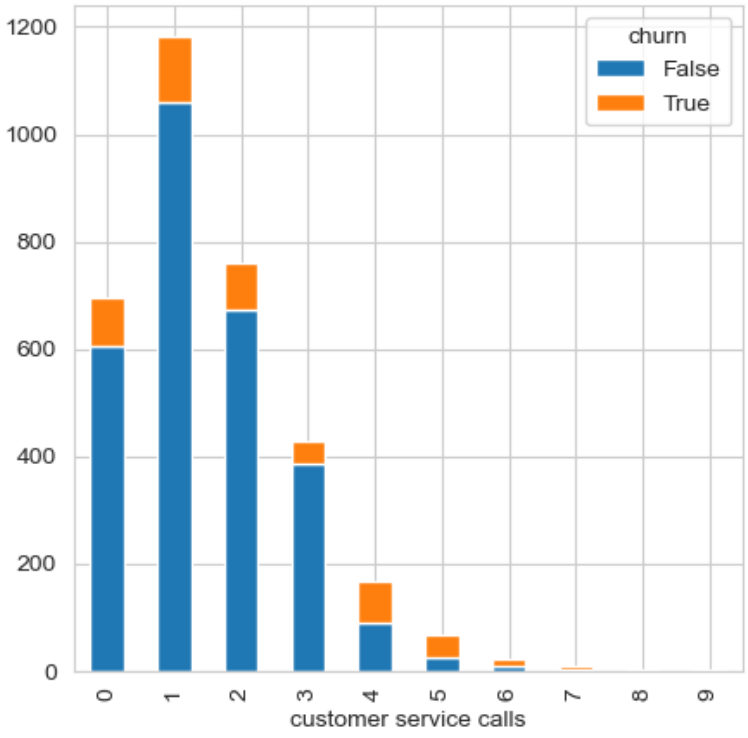

The top five most important features that determine customer churn include:

a. Customer service charge
b. Total day minutes
c. Total day charge
d. Voice mail plan
e. Area code

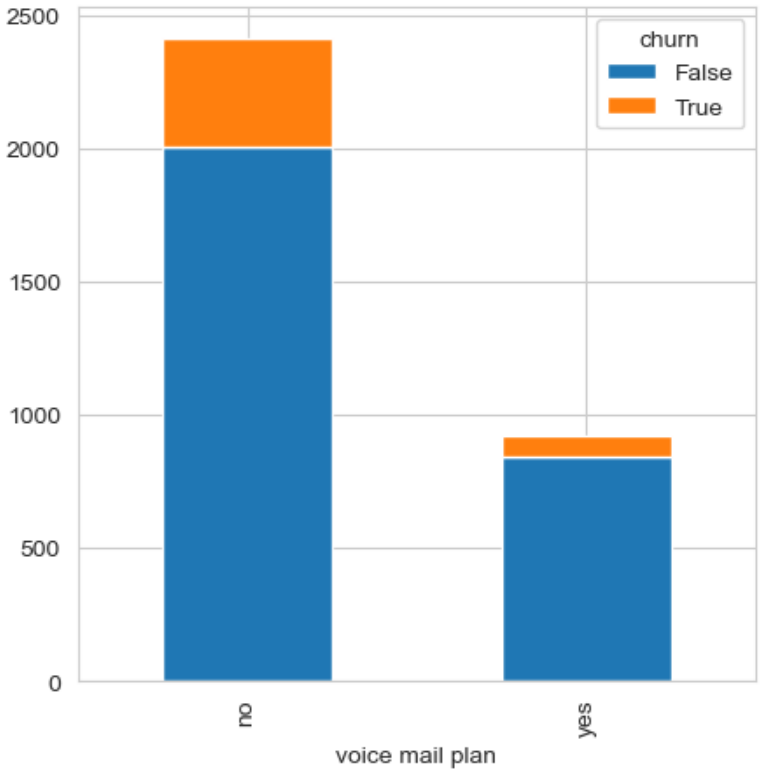
Graph of Churn against Area code




Graph of Churn against Customer Service Calls



Graph of Churn against Voice Mail Plan



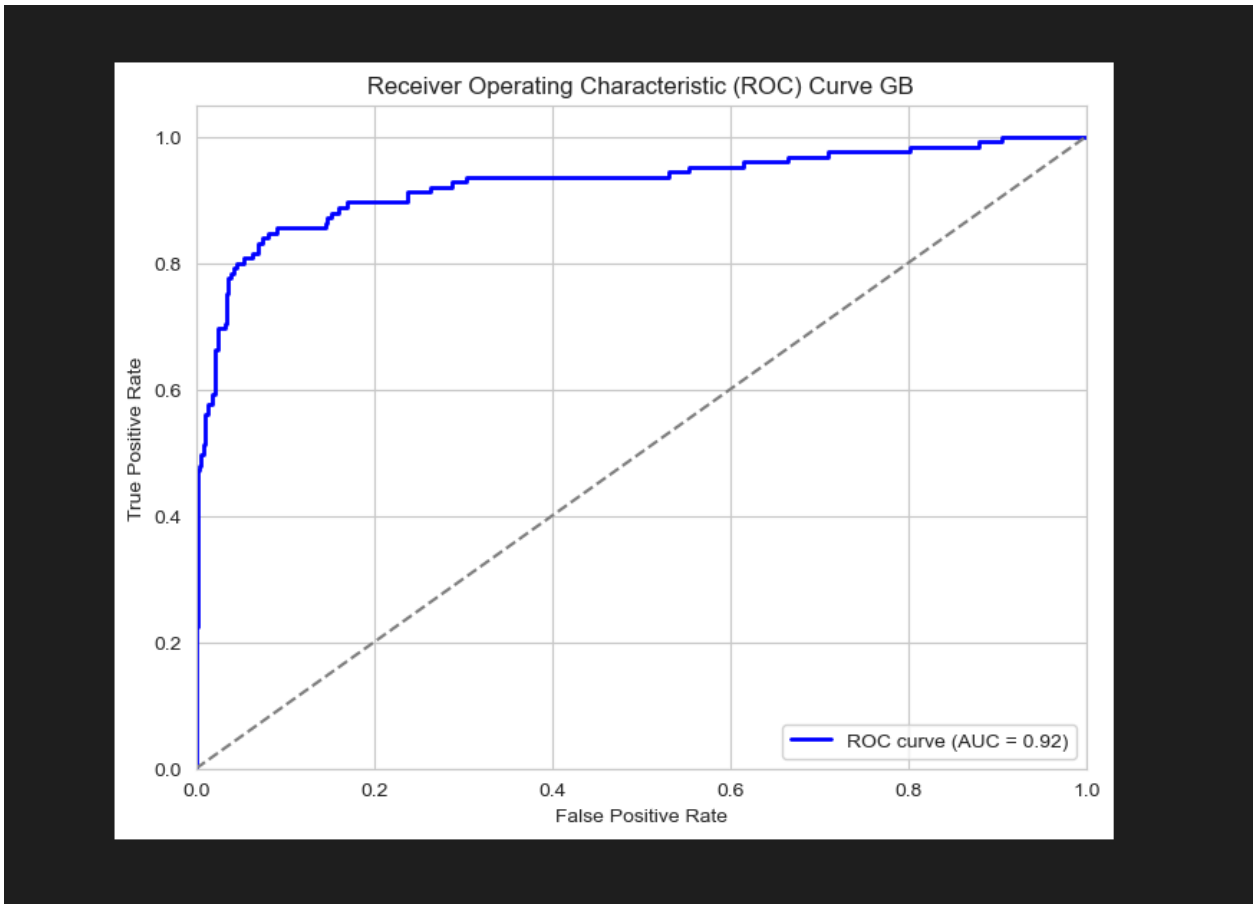

## Models' Results

| Model | Technique | Training Accuracy | Testing Accuracy | AUC |
|---|---|---|---|---|
| Logistic Regression | Imbalanced | 0.8626 | 0.8501 | 0.7609 |
| | SMOTE | 0.7274 | 0.7122 | 0.7774 |
| | SMOTE + Tuning | 0.7279 | 0.7170 | 0.7775 |
| Decision Tree | SMOTE | 1.0 | 0.8333 | 0.7800 |
| | SMOTE + Tuning | 1.0 | 0.8549 | 0.8026 |
| Random Forest | SMOTE | 1.0 | 0.9208 | 0.9132 |

| | | | | |
|---|---|---|---|---|
| | SMOTE + Tuning | 0.9915 | 0.9220 | 0.9097 |
| Gradient Boosting | SMOTE | 0.9098 | 0.9100 | 0.9091 |

The best model is gradient boosting classifier with an accuracy score of 0.9100 for test and 0.9098 for training. This model tends to be more robust to overfitting compared to other models. This model demonstrates good generalization ability.

**Visualization of ROC Curve of Gradient Boosting Classifier.**




## Limitations

- Fine Tuning Constraints: Fine-tuning hyperparameters can be time-consuming, especially for models with complex architectures or large parameter spaces.

- Computation Costs of Models with Large Parameter Spaces: Models like ensemble methods with numerous estimators require significant computational resources for training and evaluation. Training time may increase exponentially with the parameter space size, limiting scalability and applicability. High computational costs may restrict deployment in real-time or resource-constrained environments.
- Need for Comprehensive Pre-Modelling Analysis: Adequate feature engineering and exploratory data analysis are vital for identifying relevant features and understanding data distribution. Inadequate data preprocessing may lead to suboptimal performance or biased predictions. Insufficient data exploration before modelling may overlook crucial patterns, affecting interpretability and generalization.

## Recommendations

a. Customer Segmentation: Utilize the identified key features, such as total day charge, customer service call, voice mail plan, area code, and total day minutes, to segment customers based on their churn risk levels.
b. Tailor retention strategies and marketing campaigns to address the specific needs and behaviors of each segment. For example:
    - Giving discounts to area code 415 since it had the highest customer churn rate
    - Increasing marketing campaigns to area code 415
    - Improve customer service to ensure customers are adequately assisted when they make their first call
    - Include voice mail plan in the standard package
    - To provide loyalty rewards, bonus minutes, and special discounts to customers with high total day charge
    - Offer specialized plans that provide discounted rates for calls made during the day
c. Use Gradient Boosting classifier as the model of choice for forecasting. The Gradient Boosting Classifier has demonstrated strong predictive power, robustness to overfitting in churn prediction.
d. Intensive exploratory Data Analysis (EDA): Perform comprehensive exploratory data analysis to uncover hidden patterns and insights in the data. Visualizations and descriptive statistical analyses can help identify relationships between features and the target variable, guiding feature selection and modeling decisions.

## Conclusion

In conclusion, based on the model's accuracy and auc scores, Gradient Boosting classifier is the best model to predict churn of SyriaTel's customers towards strategizing, saving

costs, and prioritizing resources to increase profits. Also, customer service call, total day charge, total day minute, voice mail plan, and area codes are the most important that determine whether a customer will churn or note

## Future Work

- Explore advanced machine learning techniques, such as deep learning or ensemble methods, to further improve predictive accuracy and model performance.

### REPOSITORY STRUCTURE

• Data: Contains the raw CSV dataset ("bigml_59c28831336c6604c800002a.csv") from Syrian Tel Communication used for analysis.

• Notebooks: Includes Jupyter Notebooks detailing the entire analysis process.

• README.md: Provides an overview of the project.

• Presentation: Contains the presentation slides in pdf detailing the project findings.